African Journal of Emergency Medicine 10 (2020) S145–S149

Contents lists available at ScienceDirect

African Journal of Emergency Medicine

journal homepage: www.elsevier.com/locate/afjem

Research primer

Basic statistics: A research primer for low- and middle-income countries

Justin Kaplan[a,*], Mohammad Jalili[b], David McD. Taylor[c,d]

[a] Merck & Co Inc, North Wales, PA, United States of America
[b] Emergency Medicine Department, Tehran University of Medical Sciences, Tehran, Iran
[c] Department of Medicine, University of Melbourne, Parkville, Victoria, Australia
[d] Emergency Medicine Research, Austin Health, Heidelberg, Victoria, Australia

ARTICLE INFO

Keywords:
Publishing
Evidence-based medicine
Statistics

ABSTRACT

Statistics can be used to describe data or make inferences about populations using samples. Median values (the 50th percentile) better represent central tendency of data samples than means (averages), particularly when data have extreme values. Errors resulting from use of inferential statistics when using classical hypothesis testing include type I (finding a difference between groups when one does not exist) and type II (failure to find a true difference) errors. Confounding variables (those that vary with both the dependent variable and independent variable) may lead to spurious associations. Classical hypothesis testing and reporting only *p*-values tends to be greatly overused and overemphasized. Confidence intervals provide a range of values for a sample within a certain probability (commonly 95%). Confidence intervals can thus describe sizes of likely differences between samples, and are much more clinically useful information than only *p*-values. Before doing a study, the required sample size should be calculated to assess study feasibility. Doing so requires specification of the acceptable risk of type I and II errors and the size of the lowest clinically meaningful difference between groups.

African relevance

- Statistics is an essential tool for analysing data and establishing associations and causality.
- Several studies have reported deficiencies in clinicians' knowledge of statistics.
- A basic knowledge of statistics is necessary for authors from all settings, including limited resource settings.

The International Federation for Emergency Medicine global health research primer

This paper forms part 11 of a series of how to papers, commissioned by the International Federation for Emergency Medicine. It describes the often-challenging process of managing basic statistics. Although the authors describe the concepts eloquently, performing basic statistics often require basic statistical training. We have also included additional tips and pitfalls that are relevant to emergency medicine researchers.

Background

Statistics is an essential tool for analysing data and establishing associations and causality. Statistics helps improve the reasonableness and accuracy of inferences made in medical research and prevents errors and biases. However, statistics contributes to medical care in ways beyond research. Understanding statistics helps clinicians comprehend and appraise the empirical studies that comprise the evidence behind clinical practice [1,2]. Several studies have reported deficiencies in clinicians' knowledge of statistics, leading to misinterpretation or ignoring the statistics in published manuscripts [3,4]. Understanding statistical analysis is, therefore, required for both establishing and interpreting the evidence used to support clinical practice [5]. Statistics can be used to describe data or to make inferences, using data from a sample or samples, about a population or differences between populations. This manuscript addresses the fundamental principles of both descriptive and inferential statistics, including common statistical tests and confounders.

Types of data

Data are pieces of information that are collected, for example, through a study. Data can be numerical (quantitative) or categorical (divided into groups). It is important to recognize the type of data to select the best visualization method, an appropriate statistical analytical method, and make correct conclusions.

Quantitative data can be measured objectively. They can be:

* Corresponding author.
E-mail address: justin.kaplan@merck.com (J. Kaplan).

https://doi.org/10.1016/j.afjem.2020.06.007
Received 16 December 2019; Received in revised form 28 March 2020; Accepted 17 June 2020
Available online 11 July 2020

- Discrete: distinct and separate, non-overlapping values; determined by counting. They can only take particular values (e.g. number of abortions).
- Continuous: can be complex numbers; not restricted to arbitrarily defined separate values. They can take any value over a continuous range.

Categorical data, however, cannot be measured, but can be observed and characterized, objectively or subjectively. They can be:

- Ordinal: in rank order, but not measured. Examples include scales (e.g. Glasgow Coma Scale) and degree of pain severity (e.g. mild, moderate, severe; or a Visual Analog Scale score).
- Nominal: No ranking, just categories. Examples include gender and presence or absence of some characteristic (e.g. a risk factor).

Distributions of data

Data distribution can be understood more easily when expressed graphically. Commonly, the data of interest (e.g. counts) are plotted on a graph's vertical (Y) axis, against the data variables (values) on the horizontal (X) axis. For example, peoples' blood pressure (Y-axis) can be expressed in terms of their ages (X-axis). Data can be distributed in countless ways, but certain patterns are common. For example, data can be:

- Equally distributed across all possibilities (called a rectangular, or uniform, distribution). One example would be the birth dates of people in a population. The Y axis (the "curve") should approximate a straight horizontal line.
- Clustered around a middle range, with rates decreasing symmetrically as values deviate from the middle value, and the curve having a bell shape; this distribution can be called normal, Gaussian, or parametric. Many measured variables subject to normal biologic variability tend to have this distribution. Examples include height, weight, and body temperature. Many of the most commonly used statistical tests are designed for data that are normally (parametrically) distributed. Fig. 1a.
- Clustered asymmetrically around part of the distribution (skewed distribution): an example is household income in the USA; most people have incomes within a certain, modest range, but a very few have incomes many times greater. These data skew toward the lower incomes, an example of leftward, or positive skew. Fig. 1b.
- Clustered around more than one peak value: an example is the incidence of spontaneous pneumothorax by age—most common in young adults and the elderly, with lower rates in ages between these. Each peak is called a mode; hence this curve is called multimodal (in this case, with two modes, bimodal).

Descriptive statistics

Descriptive statistics is used to describe and summarize features of a data set. They help summarize large amounts of data and present them logically and comprehensibly.

Distributions of data depict summaries of the frequency of each value of a variable. Frequency distribution is a common way of describing a single variable. Central tendency of a distribution denotes an estimate of the "centre" (single most representative value) of a frequency distribution. The point estimate is a single value that best estimates this central tendency. The most common point estimates are the:

- Mean (average): the sum of all values in a sample divided by the number of observations
- Median (the 50th percentile): When data are ranked, half (50%) of the values are below the median and the other 50% are above it
- Mode: The single most common value

Particularly when data have extreme values, the median is more representative than the mean. For example, if almost all patients with a certain cancer survive for only a few months, but a few survive for many years, the extreme values increase the mean to a number higher than most of the patients in the sample. Thus, mean survival would be much higher than the median, more characteristic value.

Statistics can also describe the data range and dispersion. Measures include the following:

- Range: The values ranging from lowest through the highest values
- Standard deviation (SD): Indicates dispersion (how spread apart values are) *in a parametric distribution*. About 66% of values are within 1 SD either side of the mean and 95% of values are within 2 SD.

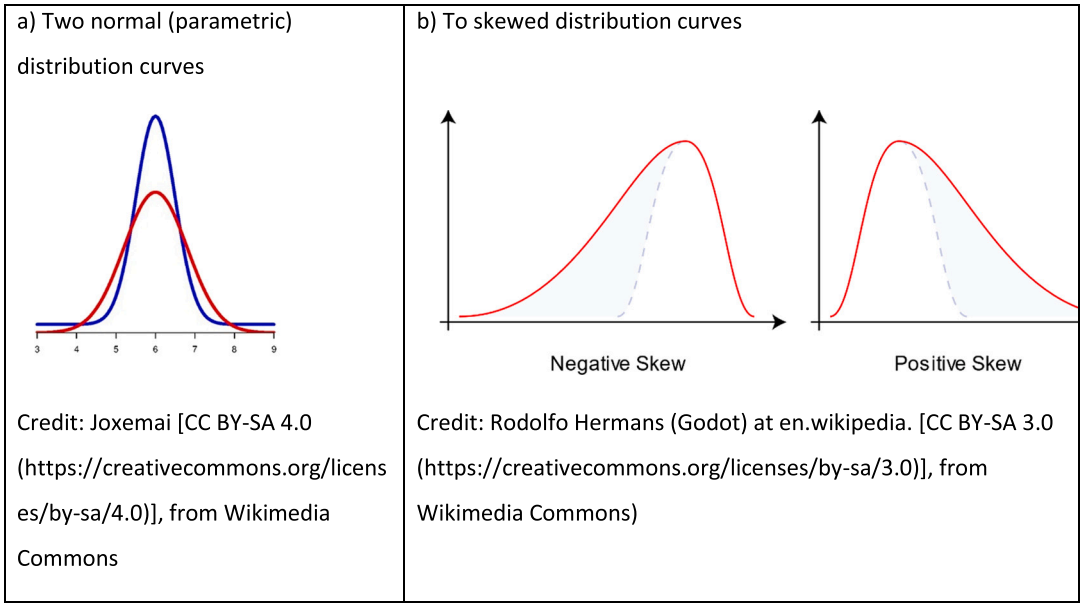

Fig. 1. Example of normal (a) and skewed (b) distribution curves.

- Interquartile range (IQR): Values ranging from the 25th through the 75th percentile in rank. IQR values are particularly useful for data not distributed parametrically.

Inferential statistics

Classical hypothesis testing

This is the most commonly used statistical method for finding differences between groups. Bayesian analysis (see reference [6]), a different approach that combines prior, known information (for example, from prior studies) with the current data, is beyond the scope of this review.

Clinical studies are often designed to identify differences in a variable (any entity that can take on different values, e.g. blood pressure or pain score) between two (or more) groups. Classical hypothesis testing aims to identify true differences in the variable of interest between these groups. True differences are differences unlikely to have resulted by random variation between the two (or more) samples representing the groups. Classical hypothesis testing is indirect and not intuitive. It starts with the 'null' hypothesis–that there are no true differences between the groups. However, the 'alternative' hypothesis is that true differences do exist (for example, that a treated group's blood pressures are different from a placebo group and thus the treatment is efficacious). This is the clinical question the study is designed to address. The *p*-value is the probability of obtaining, solely by chance, a between-group difference as large or larger than what was found. If this probability is lower than a pre-determined value designated as alpha (by convention, usually 0.05, corresponding to a 5% probability), the null hypothesis is rejected and the alternative hypothesis accepted. Therefore, alpha can be defined as the predetermined risk of committing an error of obtaining false positive results.

A type I error, analogous to a *false-positive*, is finding a difference between groups when no difference exists. This risk equals alpha, and so is typically, by convention, 5%.

A Type II error, analogous to a *false-negative*, is failure to find a true difference between groups. The chance of a type II error (called beta) is the probability of accepting the null hypothesis when the null hypothesis is false. In contrast, statistical power is the probability of finding a true difference, of a given size, between groups. A power value of 0.80 (80%) is commonly chosen, resulting in a beta value of 20%. However, many argue for a smaller beta, to maximize power and minimize the chance of falsely negative studies. Table 1 shows the relations of true and between-group differences, study findings, and errors.

For a given sample size, alpha and beta are inversely related. For example, the lower the p-value, the less likely is a (false-positive) type I error, but the more likely is a (false-negative) type II error. The only way to decrease both type I and II errors is to increase sample size, something that increases precision of estimates in statistics. Clinical criteria help inform values for alpha and beta. For example, for highly toxic interventions, a low alpha value might be preferred, to establish more conclusive evidence of efficacy. For more severe or often undertreated diseases for which treatment is benign (eg, antibiotic treatment for chlamydia salpingitis), a low beta, to minimize the chance of missing evidence of efficacy, might be preferred.

Sample size calculation

The sample size required should be calculated when planning the study. It should be based on the study's primary hypothesis and outcome variable. To calculate a sample size, the following are required:

- Alpha
- Beta
- The size of the smallest discernible difference between groups (called the effect size; see reference [7]). The effect size ideally is the minimum clinically important difference.

The required sample size increases when any of the following decreases:

- Beta (eg, using 0.05 rather than 0.20): when a higher power is desirable, the sample size needs to be larger.
- The effect size (eg, finding a difference of 5 mmHg instead of 15 mmHg): when looking for a smaller effect size, a larger sample size is required.

Sample sizes can be derived from certain tables (eg, references [8] and [9]) or calculated using statistical software. Some methods assume parametric data distributions.

Failure to calculate sample size before the study is done is an extremely common mistake. It risks carrying out studies with insufficient sample sizes, exposing subjects to the risks and inconveniences of research without the counterbalancing possibility of benefit by generating new knowledge.

Common statistical tests

In inferential statistics, researchers use statistical tests designed to determine whether differences observed between samples are due to true population differences (eg, due to an intervention) or chance. Among the myriad of tests commonly used for between-group differences in hypothesis testing are the following:

- Student's *t*-test: Differences in means between 2 groups; assumes parametric distribution
- Wilcoxon rank sum (Mann Whitney U) test: Differences in distributions between 2 groups; similar to the t-test, but does not assume a parametric distribution
- Chi-square test: Differences using qualitative data, comparing at ≥ 5 combinations of treatment and outcome. Commonly used to compare proportions.
- Fisher's exact test: Similar to chi-square, but can be used with < 5 combinations
- One-way Analysis of Variance (ANOVA): Differences in means between > 2 groups, assumes a parametric distribution and is like a *t*-test for > 2 groups
- Kruskal Wallis: Non-parametric analogue to one-way ANOVA. Like a Wilcoxon rank sum test for > 2 groups.

Confounding

Experimental studies are often designed to look for the possible effects of changing or controlling some variables (called independent or

Table 1
The relations of true and between-group differences, study findings, and errors.

	True Population Between-Group Differences	No True Population Between-Group Differences
Sample shows between-group differences	No error ("true positive")	Type I error ("false-positive"), alpha
Sample shows no between-group differences	Type II error ("false-negative"), beta	No error ("true negative")

predictor variables) on another variable (called the dependent or outcome variable). For example, one could test whether regionalization of trauma care (an independent variable) in city A has affected the trauma mortality rate (dependent variable). Confounding variables are those that vary with both the dependent variable and independent variable and are absent from the causal pathway [10]. The presence of such a confounder, unless accounted for, can cause a spurious association. For example, suppose the emergency department (ED) in city A is found to have a lower mortality rate for major trauma than that in city B. However, city B serves many nursing homes, resulting in a much older patient group, which is more susceptible to mortality from trauma. Age affects both which ED patients are likely to be in *and* the risk of mortality after trauma, *and* is not involved in a causal pathway (i.e., it does not mean that a causal relationship exists between city A and trauma mortality). Thus, age is a confounder. If confounders are ignored, the study results will falsely suggest a causal relationship.

Many techniques can adjust or control for potential confounders [11,12]. These include randomization, restriction, matching, and stratification. These methods are best incorporated into the study design. When this is not feasible or practical, researchers need to apply a statistical "correction" method during analysis to adjust for potentially confounding effects. Regression is one such core statistical technique. Regression can be used to analyse quantitative data (called linear regression), qualitative data (logistic regression), or time to an event, such as survival data (Cox regression). Logistic regression is used when the dependent variable is categorical and linear regression is used when the dependent variable is continuous.

Interpreting and reporting results

p-Values

For *interpretation* of study results, classical hypothesis testing and *p*-values tend to be greatly overused and overemphasized. Their use has been criticized for being non-intuitive, and thus prone to misinterpretation. They also provide too little useful information (see reference [13]). A low *p*-value, indicating differences between groups, results from one of the following: true causation (a real cause-and-effect relationship), chance (a source of random error), confounding, or bias (systematic error resulting from flaws in data sampling or measurement). A low p-value is evidence of the absence of chance - only one of the factors above.

Also, *p*-values, because they say nothing about the size of the difference, provide no information about *clinical* significance. Point estimates and confidence intervals provide much more information, particularly clinically useful information.

Confidence intervals

Confidence intervals provide a range of values within a certain probability. For example, if sampled repeatedly, the mean or median would be within the range of the 95% confidence interval 95% of the time. This is often oversimplified to be the range within which the true value lies with 95% probability. Confidence intervals can be used to describe single group samples or differences between groups. If a between-group confidence interval does not include a zero difference, that between-group confidence interval is statistically significant. For example, if the 95% confidence interval of the difference between the effect of anti-hypertensive drug A and drug B on blood pressure was 1 to 5 mmHg, that difference is statistically significant. That is, the effect of drug A differs from that of drug B over the entire range of values in the confidence interval; therefore, the likelihood that this observed difference is due to chance is outside the 95% range (and thus is $< 5\%$).

Confidence intervals give much more clinically useful information than *p*-values and are generally preferred by statisticians and reviewers. Confidence intervals show how precise point estimates are. (Larger sample sizes result in narrower, more precise, confidence intervals.) Confidence intervals also can show how likely differences are to be clinically significant (see reference [3]). In the example above, although differences are statistically significant, the 95% confidence intervals are only 1 to 5 mmHg, which are probably not considered clinically significant. In contrast, if the interval was found to be 20 to 30 mmHg, differences would be considered clinically significant. If the interval was 1 to 20 mmHg, the true difference, although statistically significant, ranges from clinically insignificant (1 mmHg) to clinically significant (20 mmHg). Large confidence intervals, often with clinically and sometimes statistically indeterminate results, commonly result because sample sizes are too small.

Tips on this topic

- Consult a statistician during the design phase of the study. Little can be done to fix design problems after the study is completed.
- Calculate the required sample size and consider study designs that minimize sample size.
- Provide more information than *p*-values (eg, point estimates, measures of variability) when reporting results, even in abstracts.
- Use confidence intervals whenever possible.
- Anticipate possible confounders and adjust for them in study design and/or analysis.

Pitfalls to avoid

- Overemphasizing the *p* value; it says nothing about clinical significance and is widely misinterpreted.
- Failure to calculate sample size before beginning a study.
- Failure to define types of variables, which leads to wrong choice of statistical tests.

Additional relevant issues

Among the statistical topics important in emergency medicine that are beyond the scope of this review are the following:

- Regression, a statistical method that quantifies the strength of association between an outcome variable (eg, morbidity, mortality) and one or more independent variables (eg, age, Acute Physiology and Chronic Health Evaluation [APACHE] score). Regression can adjust for confounding variables (eg age, which can confound the association between variables predicting various poor clinical outcomes).
- Multiple comparisons: Because every hypothesis tested has a chance of being positive solely by chance, the more hypotheses that are tested within a single analysis, the greater the chance that at least one will be positive by chance. Statistical analyses should be modified to account for multiple comparisons; including subgroup analyses (see reference [7]).
- Minimizing sample sizes: Minimizing the required sample is high priority if resources, including subjects, are limited. Certain study designs reduce the required sample size (see[1] 15 Ways to Reduce Sample Size in Clinical Trials).

Annotated bibliography

These readings cover relevant topics in greater depth than possible in this manuscript. They are worth reading in their entirety. Reference number 1, in particular, is a lucid explanation of essential content.

1. Braitman LE. Confidence intervals assess both clinical significance and statistical significance. Ann Intern Med. 1991;114(6):515–517

[1] https://blog.statsols.com/15-ways-to-reduce-sample-size-in-clinical-trials.

2. Dmitrienko A, D'agostino RB. Multiplicity Considerations in Clinical Trials. NEJM. 2018;378(22):2115–2122

Author contribution

Authors contributed as follow to the conception or design of the work; the acquisition, analysis, or interpretation of data for the work; and drafting the work or revising it critically for important intellectual content: JK contributed 75%, MJ contributed 15% and DT contributed 10%. All authors approved the version to be published and agreed to be accountable for all aspects of the work.

Declaration of competing interest

The authors declare no conflict of interest.

References

[1] Aggarwal R. Statistical literacy for healthcare professionals: why is it important? Ann Card Anaesth 2018;21(4):349–50. https://doi.org/10.4103/aca.ACA_177_18.
[2] Barkan H. Statistics in clinical research: important considerations. Ann Card Anaesth 2015;18(1):74–82. https://doi.org/10.4103/0971-9784.148325.
[3] Best AM, Laskin DM. Oral and maxillofacial surgery residents have poor understanding of biostatistics. J Oral Maxillofac Surg 2013;71:227–34.
[4] Bookstaver PB, Miller AD, Felder TM, Tice DL, Norris LB, Sutton SS. Assessing pharmacy residents' knowledge of biostatistics and research study design. Ann Pharmacother 2012;46:991–9.
[5] Hack JB, Bakhtiari P, O'Brien K. Emergency medicine residents and statistics: what is the confidence? J Emerg Med 2009;37:313–8.
[6] Quintana M, Viele K, Lewis RJ. Bayesian analysis; using information to interpret the results of clinical trials. JAMA 2017;318(16):1605–6.
[7] Cook JA, Julious SA, Sones W, et al. DELTA2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomized controlled trial. BMJ 2018;363:k3750C.
[8] Fleiss JL. Statistical methods for rates and proportions. 3rd ed. Hoboken, New Jersey: John Wiley & Sons; 2003.
[9] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, New Jersey: Lawrence Erlbaum Associates; 1988.
[10] VanderWeele TJ, Shpitser I. On the definition of a confounder. Ann Stat 2013;41:196–220.
[11] Vetter TR, Mascha EJ. Bias, confounding, and interaction: lions and tigers, and bears, oh my!. Anesth Analg 2017;125(3):1042–8. Sep.
[12] Pourhoseingholi MA, Baghestani AR, Vahedi M. How to control confounding effects by statistical analysis. Gastroenterol Hepatol Bed Bench 2012;5(2):79–83.
[13] Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol 2016;31:337–50.